A Textual Information Detection and Elimination System for Secure Medical Image Distribution *

James Ze Wang, M.S., Michel Bilello, Ph.D., Gio Wiederhold, Ph.D.
School of Medicine, Stanford University, Stanford, CA 94305

Background. As the demand for greater accessibility to health care information grows, medical institutions are being urged to make information available to legitimate external parties in a timely fashion (*e.g.*, on-line) while protecting the privacy of patient data. It is therefore crucial that health care institutions be provided with on-line tools that allow them to disseminate medical information without compromising data privacy. We present an algorithm that strips textual information (including identifying information) from medical images such as digitized x-rays and CT scans. The resulting processed images can then be made available to medical researchers, physicians, and other legitimate users. Such a tool could be used by health care institutions and other repositories of medical images as part of a data security system.

Algorithm. In the TIDE (Textual Information Detection and Elimination system for secure medical image distribution) project, we developed an efficient and accurate algorithm to distinguish areas with and without textual information in medical images. Because variations in the diagonal directions can be found in almost all Roman characters or Arabic numbers, we use Daubechies' wavelets and post-processing techniques to detect the high frequency variation in the diagonal direction that is indicative of text.

We apply a 1-level fast wavelet transform (FWT) with Daubechies' Symlet-8 wavelet or Daubechies-8 wavelet to each image. Then we extract and post-process the lower right-hand corner of the transform matrix, where the diagonal directional high frequency information is located, to obtain a mask containing only the areas with textual information. Once such a mask is computed, we may apply it to the original image to eliminate the areas with textual data.

Our design has several immediate advantages.

1. Unlike traditional approaches, such as the neural network, our algorithm does not depend on the actual font and style of the text in the medical image. Preliminary experiments indicate that the algorithm is capable of handling images with superimposed hand-written text.
2. We used Daubechies' wavelets rather than a traditional edge detector to capture the high frequency information in the images. This reduced the dependence of the results on the quality or the sharpness of the images.
3. It does not rely on the color of the image or the text. It also has minimum dependence on the contrast between text and background objects.

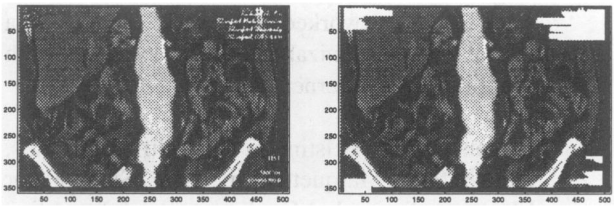

Figure 1: **A medical image with hand-written text processed by the TIDE system.** Text Information simulated.

Results. This algorithm has been implemented on a Sparc-20 workstation. We have tested about 30 medical images of different types, collected from different sources. It takes about 10 seconds of CPU time to process each medical image of size 512 × 512. Besides the fast speed, the algorithm has achieved remarkable accuracy. It successfully detected and eliminated all of the critical textual information within the medical images, many with superimposed text.

Conclusions and Future Work. We have demonstrated an efficient wavelet-based textual information detection and elimination system for secure medical image distribution.

We are working on applying this technique to large number of real-world medical images. We are also trying to improve this technique so that only texts related to patients' private information, e.g. patient name or patient identification number, are eliminated.

[*] *Correspondence to:* wangz@cs.stanford.edu

1091—8280/97/$5.00 © 1997 AMIA, Inc.